On Metrizing Vague Convergence of Random Measures with Applications on Bayesian Nonparametric Models

Luai Al-Labadi*

Abstract

This paper deals with studying vague convergence of random measures of the form $\mu_n = \sum_{i=1}^{n} p_{i,n}\delta_{\theta_i}$, where $(\theta_i)_{1\leq i\leq n}$ is a sequence of independent and identically distributed random variables with common distribution Π, $(p_{i,n})_{1\leq i\leq n}$ are random variables chosen according to certain procedures and are independent of $(\theta_i)_{i\geq 1}$ and δ_{θ_i} denotes the Dirac measure at θ_i. We show that μ_n converges vaguely to $\mu = \sum_{i=1}^{\infty} p_i\delta_{\theta_i}$ if and only if $\mu_n^{(k)} = \sum_{i=1}^{k} p_{i,n}\delta_{\theta_i}$ converges vaguely to $\mu^{(k)} = \sum_{i=1}^{k} p_i\delta_{\theta_i}$ for all k fixed. The limiting process μ plays a central role in many areas in statistics, including Bayesian nonparametric models. A finite approximation of the beta process is derived from the application of this result. A simulated example is incorporated, in which the proposed approach exhibits an excellent performance over several existing algorithms.

KEY WORDS: Beta process, Nonparametric Bayesian statistics, Point Processes, Random measures, Vague convergence.

MSC 2000 Primary 62F15, 60G57; Secondary 28A33.

**Address for correspondence:* Luai Al-Labadi, Department of Mathematical & Computational Sciences, University of Toronto Mississauga, 3359 Mississauga Road, Deerfield Hall, Room 3015, Mississauga, Ontario M5S 3G3, Canada. E-mail: luai.allabadi@utoronto.ca.

1 Introduction

The primary objective of this paper is to study the vague convergence of a particular class of random measures of the form

$$\mu_n = \sum_{i=1}^{n} p_{i,n}\delta_{\theta_i}, \tag{1}$$

where $(\theta_i)_{i\geq 1}$ is a sequence of independent and identically distributed (i.i.d.) random variables with common distribution Π, $(p_{i,n})_{1\leq i\leq n}$ are random variables chosen according to certain procedures and are independent of $(\theta_i)_{i\geq 1}$, and δ_{θ_i} denotes the Dirac measure at θ_i. In particular, we show that μ_n converges vaguely to $\mu = \sum_{i=1}^{\infty} p_i\delta_{\theta_i}$ if and only if $\mu_n^{(k)} = \sum_{i=1}^{k} p_{i,n}\delta_{\theta_i}$ converges vaguely to $\mu^{(k)} = \sum_{i=1}^{k} p_i\delta_{\theta_i}$ for all k fixed. The limiting process μ has played a central role in some area in statistics, including Bayesian nonparametric models. Interesting examples of μ include, among others, the Dirichlet process (Ferguson, 1973), the beta process (Hjort, 1990), the beta-Stacy process (Walker and Muliere, 1997), the two parameter Poisson-Dirichlet process (Pitman and Yor, 1997) and the normalized inverse-Gaussian process (Lijoi, Mena and Prünster, 2005). For a recent summary of the Bayesian nonparametric priors and their applications, please refer to the book of Phadia (2013) and Müller, Quintana, Jara and Hanson (2015). A motivating application of the aforementioned result involves the derivation of a finite sum representation that converges vaguely to the Wolpert and Ickstadt (1998) representation of the beta process.

This paper is organized as follows. Section 2 discusses metrizing the vague convergence of random measures of the form (1). It also develops a criterion for which μ_n converges vaguely to μ. In Section 3, a finite sum approximation of the beta process is derived. In Section 4, an example comparing the performance of

the new approximation to other existing approximations is presented. Section 5 ends with a brief summary of the results.

2 Metrizing Vague Convergence of Random Measures

The material developed in this section can be seen as a convenient adaptation of the work of Grandell (1977). Let μ be a measure on $\mathbb{R}$ such that its distribution function $\mu(t) = \mu((-\infty, t])$, $t \in \mathbb{R}$, is finite for finite t. The same notation will be used for the measure and its distribution function. We assume that $\mu(-\infty) = 0$. The $\mu-$measure of the interval $(a, b]$ is denoted by $\mu((a, b])$ or in terms of the distribution function by $\mu(b) - \mu(a)$. The set of measures μ on $\mathbb{R}$ such that $\mu(-\infty) = 0$ and $\mu(t) < \infty$ for all $t \in \mathbb{R}$ is denoted by $\mathcal{M}$. We shall now define vague convergence on $\mathcal{M}$.

Definition 1 *Let $\mu, \mu_1, \mu_2, \ldots \in \mathcal{M}$, be given. We say that μ_n converges vaguely to μ and write that $\mu_n \xrightarrow{v} \mu$ if $\mu_n(t) \to \mu(t)$ for all $t \in \mathbb{R}$ such that μ is continuous at t.*

Let $C_K^+(\mathbb{R})$ be the set of all nonnegative continuous real valued functions with compact support defined on $\mathbb{R}$. In $C_K^+(\mathbb{R})$ all functions are bounded and for each $f \in C_K^+(\mathbb{R})$ there exists to a number x_f such that $f(x) = 0$ for all $x > x_f$. For the proof of the next theorem consult, for example, Grandell (1977).

Theorem 2 *Let $\mu, \mu_1, \mu_2, \ldots \in \mathcal{M}$ be given. Then, as $n \to \infty$, $\mu_n \xrightarrow{v} \mu$ if and only if*

$$\int_{-\infty}^{\infty} f(x)\mu_n(dx) \to \int_{-\infty}^{\infty} f(x)\mu(dx)$$

for all $f \in C_K^+(\mathbb{R})$.

To define a metric, corresponding to vague convergence, in the set $\mathcal{M}$, we first need to define a metric in the following set:

$$\mathcal{M}^{(k)} = \left\{\mu \in \mathcal{M} : \mu(x) = \mu(\theta_{(k)}) \text{ for } x \geq \theta_{(k)})\right\},$$

where $\theta_{(k)} = \max_{1\leq i\leq k} \theta_i$. Here, $\theta_i \overset{\text{i.i.d}}{\sim} \Pi$, where Π is a fixed continuous probability measure on $\mathbb{R}$. In other words, the set $\mathcal{M}^{(k)}$ is the set of those measures in $\mathcal{M}$ with total mass on the interval $(-\infty, \theta_{(k)}]$. It is worth mentioning that the set $\mathcal{M}^{(k)}$ considered in this section is different from that defined in Grandell (1977). Grandell (1977) used the following set:

$$\mathcal{M}^{(k)}_{\text{Grandell}} = \{\mu \in \mathcal{M} : \mu(x) = \mu(k) \text{ for } x \geq k\}.$$

Clearly, the random measure μ_k defined in (1) puts all its mass in $(-\infty, \theta(k)]$. Thus, μ_k belongs to $\mathcal{M}^{(k)}$ but not to $\mathcal{M}^{(k)}_{\text{Grandell}}$.

For μ_1 and $\mu_2 \in \mathcal{M}^{(k)}$, the Lévy metric is defined by

$$d_L(\mu_1, \mu_2) = \inf\{h \geq 0 : \mu_1(x-h) - h \leq \mu_2(x) \leq \mu_1(x+h) + h, \forall x \in \mathbb{R}\}.$$

The next Lemma deals with some properties of d_L. The proof is very similar to the proof of Lemma 1 of Grandell (1977). Thus, the proof is omitted.

Lemma 3 *d_L is a metric in $\mathcal{M}^{(k)}$, i.e. for all $\mu_1, \mu_2, \mu_3 \in \mathcal{M}^{(k)}$ we have*

1. $d_L(\mu_1, \mu_2) = 0$ *if and only if* $\mu_1 = \mu_2$.
2. $d_L(\mu_1, \mu_2) = d_L(\mu_2, \mu_1)$.
3. $d_L(\mu_1, \mu_3) \leq d_L(\mu_1, \mu_2) + d_L(\mu_2, \mu_3)$

The proof of the next lemma follows by imitating the proof of Lemma 2 of Grandell (1977) with n and k are replaced by $\theta_{(k)}$ and n, respectively.

Lemma 4 *d_L metrizes vague convergence in $\mathcal{M}^{(k)}$, i.e. $\mu_n \xrightarrow{v} \mu$ (as $n \to \infty$) if and only if $d_L(\mu_n, \mu) \to 0$ (as $n \to \infty$) for $\mu, \mu_1, \mu_2, \ldots \in \mathcal{M}^{(k)}$.*

In Lemma 4, we have shown that the Lévy metric metrizes vague convergence in $\mathcal{M}^{(k)}$. We will use this to prove a similar result in $\mathcal{M}$. As in Grandell (1977), the idea is to associate to each $\mu \in \mathcal{M}$ a vector $\left(\mu^{(1)}, \mu^{(2)}, \ldots\right)$, where $\mu^{(k)} \in \mathcal{M}^{(k)}$ and where componentwise convergence is equivalent to convergence. To do this, choose $f_1, f_2, \cdots \in C_K^+(\mathbb{R})$ such that

$$f_k(t) = \begin{cases} 1 & t < \theta_{(k-1)} \\ \theta_{(k)} - t & \theta_{(k-1)} \leq t < \theta_{(k)} \\ 0 & t \geq \theta_{(k)} \end{cases}$$

Define $\mu^{(k)}(t) = \int_{-\infty}^{t} f_k(x)\mu(dx)$. Clearly, the mapping $\mu \curvearrowright \left(\mu_1^{(k)}, \mu_2^{(k)}, \cdots\right)$ is one to one. Define, for μ_1 and $\mu_2 \in \mathcal{M}$,

$$d\left(\mu_1, \mu_2\right) = \sum_{k=1}^{\infty} \frac{d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right)}{2^k\left(1 + d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right)\right)}.$$

Theorem 5 *d is a metric on $\mathcal{M}$, i.e.*

1. *$d(\mu_1, \mu_1) = 0$ if and only if $\mu = \mu_2$*
2. *$d(\mu_1, \mu_2) = d(\mu_2, \mu_1)$.*
3. *$d(\mu_1, \mu_3) \leq d(\mu_1, \mu_2) + d(\mu_2, \mu_3)$ for all $\mu_1, \mu_2, \mu_3 \in \mathcal{M}$.*

Proof. (i) $d(\mu_1, \mu_1) = 0$ if and only if $d_L(\mu_1^{(k)}, \mu_2^{(k)})$ if and only if $\mu_1^{(k)} = \mu_2^{(k)}$ if and only if $\mu_1 = \mu_2$.

(ii) $d(\mu_1, \mu_2) = d(\mu_2, \mu_1)$ follows directly since $d_L(\mu_1^{(k)}, \mu_2^{(k)}) = d_L(\mu_2^{(k)}, \mu_1^{(k)})$.

(iii) Since $d_L(\mu_1^{(k)}, \mu_3^{(k)}) \le d_L(\mu_1^{(k)}, \mu_2^{(k)}) + d_L(\mu_2^{(k)}, \mu_3^{(k)})$ for all $\mu_1^{(k)}, \mu_2^{(k)}, \mu_3^{(k)} \in \mathcal{M}^{(k)}$, we have

$$\begin{aligned}
\frac{d_L\left(\mu_1^{(k)}, \mu_3^{(k)}\right)}{1 + d_L\left(\mu_1^{(k)}, \mu_3^{(k)}\right)} &= 1 - \frac{1}{1 + d_L\left(\mu_1^{(k)}, \mu_3^{(k)}\right)} \\
&\le 1 - \frac{1}{1 + d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right) + d_L\left(\mu_2^{(k)}, \mu_3^{(k)}\right)} \\
&= \frac{d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right) + d_L\left(\mu_2^{(k)}, \mu_3^{(k)}\right)}{1 + d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right) + d_L\left(\mu_2^{(k)}, \mu_3^{(k)}\right)} \\
&= \frac{d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right)}{1 + d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right) + d_L\left(\mu_2^{(k)}, \mu_3^{(k)}\right)} \\
&\quad + \frac{d_L\left(\mu_2^{(k)}, \mu_3^{(k)}\right)}{1 + d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right) + d_L\left(\mu_2^{(k)}, \mu_3^{(k)}\right)} \\
&\le \frac{d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right)}{1 + d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right)} + \frac{d_L\left(\mu_2^{(k)}, \mu_3(k)\right)}{1 + d_L\left(\mu_2^{(k)}, \mu_3^{(k)}\right)},
\end{aligned}$$

Thus,

$$\begin{aligned}
d\left(\mu_1, \mu_3\right) &= \sum_{k=1}^{\infty} \frac{d_L\left(\mu_1^{(k)}, \mu_3^{(k)}\right)}{2^k\left(1 + d_L\left(\mu_1^{(k)}, \mu_3^{(k)}\right)\right)} \\
&\le \sum_{k=1}^{\infty} \frac{d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right)}{2^k\left(1 + d_L\left(\mu_1^{(k)}, \mu_2^{(k)}\right)\right)} + \sum_{k=1}^{\infty} \frac{d_L\left(\mu_2^{(k)}, \mu_3^{(k)}\right)}{2^k\left(1 + d_L\left(\mu_2^{(k)}, \mu_3^{(k)}\right)\right)} \\
&= d\left(\mu_1, \mu_2\right) + d\left(\mu_2, \mu_3\right).
\end{aligned}$$

This completes the proof of the lemma. ■

Theorem 6 *d metrizes vague convergence in $\mathcal{M}$.*

Proof. Let $\mu, \mu_1, \mu_2, \ldots \in \mathcal{M}$ be given. It follows from the definition of d that $d(\mu_n, \mu) \to 0$ if and only if $d_L(\mu_n^{(k)}, \mu^{(k)}) \to 0$ (as $n \to \infty$) for all k. By Lemma

4, this holds if and only if $\mu_n^{(k)} \xrightarrow{v} \mu^{(k)}$ for all k. Thus, it is enough to prove that $\mu_n \xrightarrow{v} \mu$ if and only if $\mu_n^{(k)} \xrightarrow{v} \mu^{(k)}$ for all k.

By Theorem 2, $\mu_n \xrightarrow{v} \mu$ if and only if $\int f(x)\mu_n(dx) \to \int f(x)\mu(dx)$ for all $f \in C_K^+(\mathbb{R})$. Since $ff_k \in C_K^+(\mathbb{R})$ for all k and all $f \in C_K^+(\mathbb{R})$, it follows that $\mu_n \xrightarrow{v} \mu$ implies that $\mu_n^{(k)} \xrightarrow{v} \mu^{(k)}$ for all k.

Conversely, if $\mu_n^{(k)} \xrightarrow{v} \mu^{(k)}$ for all k we can for each $f \in C_K^+(\mathbb{R})$ choose k so that $ff_k = f$. Thus, we have

$$\begin{aligned} \int f(x)\mu_n(dx) &= \int f(x)f_k(x)\mu_n(dx) = \int f(x)\mu_n^{(k)}(dx) \\ &\to \int f(x)\mu^{(k)}(dx) = \int f(x)\mu(dx).T \end{aligned}$$

Thus, $\mu_n \xrightarrow{v} \mu$. ■

The proof of Theorem 3 reveals the following interesting result.

Corollary 7 *Let $\mu, \mu_1, \mu_2, \ldots \in \mathcal{M}$ and $\mu^{(k)}, \mu_1^{(k)}, \mu_2^{(k)}, \ldots \in \mathcal{M}^{(k)}$. Then $\mu_n \xrightarrow{v} \mu$ (as $n \to \infty$) if and only if $\mu_n^{(k)} \xrightarrow{v} \mu^{(k)}$ (as $n \to \infty$) for all k fixed.*

3 Applications on Bayesian Nonparametric Priors

An interesting application of Corollary 7 comprises deriving a finite sum representation that converge vaguely to the Wolpert and Ickstadt (1998) representation of the beta process (Hjort, 1990). Finite sum approximations for the Dirichlet process, beta-Stacy process, normalized inverse-Gaussian process and two-parameter Poisson-Dirichlet process were derived, respectively, in Zarepour and Al-Labadi (2012) and Al-Labadi and Zarepour (2013a,b; 2014a,b).

There are two common techniques to writing a series representation for any Lévy process having no Gaussian component. The first comes from Ferguson

and Klass (1972). The second technique is from Wolpert and Ickstadt (1998). A brief discussion of the two methods is described in Appendix A of this paper. An interesting comparison between the two representations from the computational point of view was addressed in Al-Labadi and Zarepour (2013a). Here, it is pointed out that the representation of Wolpert and Ickstadt is more appropriate for dealing with nonhomogeneous processes (i.e., the Lévy measure in (2) depends on t). Conversely, for homogeneous processes (i.e., the Lévy measure is independent of t), the two approaches are equivalent.

Let A_0 be a continuous cumulative hazard function and $c(t)_{t\geq 0}$ be a piecewise continuous, nonnegative function. Following Hjort (1990), the beta process A, written $A \sim BP\left(c(\cdot), A_0(\cdot)\right)$, is the completely random measure with Lévy measure

$$L_t(ds) = \left[\int_0^t c(z)s^{-1}(1-s)^{c(z)-1}dA_0(z)\right]ds, \quad \text{for } t \geq 0, \quad 0 < s < 1. \tag{2}$$

By (6) and (7), since no closed form for the inverse of the Lévy measure (2) exists, the simulation of the beta process based on series representations is very complex and may be difficult to apply in practice for many users. The next theorem outlines a remedy to this problem. Note that, when $c(t) = c$ for all t (i.e. the homogenous case), Al-Labadi and Zarepour (2015) derived a finite sum approximation and showed that it converges almost surely to the representation of Ferguson and Klass (1972) of the beta process. More details about interesting properties of the beta process when $c(t) = c$ for all t are discussed Al-Labadi and Abdelrazeq (2016).

Theorem 8 *Let $(\theta_i)_{i\geq 1}$ be i.i.d. random variables with common distribution Π and $\Gamma_i = E_1 + \cdots + E_i$, where $(E_i)_{i\geq 1}$ are i.i.d. with exponential distribution with mean 1, independent of $(\theta_i)_{i\geq 1}$. Let $A \sim BP\left(c(\cdot), A_0(\cdot)\right)$ on $[0, t_0]$, where $t_0 > 0$ is fixed. We assume that A_0 is continuous with $A_0(t_0) < \infty$. Let*

$\Pi(dz) = \eta(dz)/A_0(t_0)$, *where* $\eta([0,t]) = A_0(t)$.

$$L_{n,\theta}(x) \quad = \quad \frac{\Gamma\left(c(\theta)\right)}{\Gamma(c(\theta)/n)\Gamma\left(c(\theta)-c(\theta)/n\right)} \int_x^1 s^{c(\theta)/n-1}\left(1-s\right)^{c(\theta)(1-1/n)-1} ds. \quad (3)$$

and

$$M_z(x) = A_0(t_0)c(z)\int_x^1 s^{-1}(1-s)^{c(z)-1}ds.$$

Then, as $n \to \infty$,

$$A_n(t) = \sum_{i=1}^{n} L_{n,\theta_i}^{-1}\left(\frac{\Gamma_i}{A_0(t_0)n}\right)\delta_{\theta_i} \xrightarrow{v} A(t) = \sum_{i=1}^{\infty} M_{\theta_i}^{-1}\left(\Gamma_i\right)\delta_{\theta_i}. \quad (4)$$

Proof. First we show that, for any $x \in (0,1)$,

$$nA_0(t_0)L_{n,\theta}(x) \to M_\theta(x). \quad (5)$$

Note that, for any $x > 0$,

$$\Gamma(x) = \frac{\Gamma(x+1)}{x}.$$

With $x = c(\theta)/n$ we obtain

$$\frac{n}{\Gamma(c(\theta)/n)} = \frac{c(\theta)}{\Gamma(c(\theta)/n+1)}.$$

Since $\Gamma(x)$ is a continuous function,

$$\frac{n}{\Gamma(c(\theta)/n)} \times \frac{\Gamma\left(c(\theta)\right)}{\Gamma\left(c(\theta)-c(\theta)/n\right)} \to c(\theta).$$

Clearly, the integrand in the right hand side of (3) converges to $s^{-1}(1-s)^{c(\theta)-1}$. To apply the dominated convergence theorem, we need to show that this integrand is dominated by an integrable function. Since $x < s < 1$, we have

$s^{-1} < x^{-1}$ and $s^{c(\theta)/n} < 1$. This implies that $s^{c(\theta)/n-1} < x^{-1}$. Therefore, the integrand is bounded above by the integrable function $x^{-1}(1-s)^{c(\theta)(1-1/n)-1}$. Thus, by the dominated convergence theorem, we get (5). Since that the left hand side of (5) is a sequence of a continuous monotone functions converging to a monotone function for every $x > 0$. This is equivalent to the convergence of their inverse function to the inverse function of the right hand side (de Haan & Ferreira, 2006, page 5). Thus, as $n \to \infty$,

$$L_{n,\theta_i}^{-1}\left(\frac{\Gamma_i}{nA_0(t_0)}\right) \xrightarrow{v} M_{\theta_i}^{-1}(\Gamma_i),$$

To complete the proof of the theorem, we apply Corollary 7 with

$$\mu_n^{(k)} = A_n^{(k)} = \sum_{i=1}^{k} L_{n,\theta_i}^{-1}\left(\frac{\Gamma_i}{A_0(t_0)n}\right)\delta_{\theta_{(i)}}$$

and

$$\mu^{(k)} = A^{(k)} = \sum_{i=1}^{k} M_{\theta_i}^{-1}(\Gamma_i)\,\delta_{\theta_{(i)}}.$$

Clearly, both $A_n^{(k)}$ and $A^{(k)}$ belong to $\mathcal{M}^{(k)}$. Since, for all k fixed,

$$A_n^{(k)} \xrightarrow{v} A^{(k)},$$

as $n \to \infty$, we get (4). This completes the proof of the theorem. ■

Note that, $L_{n,\theta_i}^{-1}(p)$ is the $1-p$-th quantile of $beta\,(c(\theta_i)/n, c(\theta_i)(1-1/n))$ distribution. This provides the following algorithm.

1. Fix a relatively large positive integer n.

2. For $i = 1, \ldots, n$, generate $\theta_i \overset{\text{i.i.d.}}{\sim} \Pi$, where $\Pi(dz) = \eta(dz)/A_0(t_0)$ and $\eta([0,t]) = A_0(t)$.

3. For $i = 1, \ldots, n+1$, generate $E_i \overset{\text{i.i.d.}}{\sim} exponential(1)$ such that $(E_i)_{1\le i\le n+1}$

and $(\theta_i)_{1\leq i\leq n}$ are independent.

4. For $i = 1, \ldots, n+1$, compute $\Gamma_i = E_1 + \cdots + E_i$.

5. For $i = 1, \ldots, n$, compute $L^{-1}_{n,\theta_i}\left(\Gamma_i / \left(A_0(t_0)n\right)\right)$.

6. Use (4) to obtain an approximate value of $A \sim BP(c(\cdot), A_0(\cdot))$.

Note that, it is possible to extend Theorem 8 to derive an approximation of the beta-Dirichlet process (Kim, James and Weibbach, 2012), a nonparametric prior for the cumulative intensity functions of a Markov process. Specifically, as $n \to \infty$,

$$B_n(t) = \sum_{i=1}^{n} V(\theta_i) L^{-1}_{n,\theta_i}\left(\frac{\Gamma_i}{A_0(t_0)n}\right)\delta_{\theta_i} \xrightarrow{v} B(t) = \sum_{i=1}^{\infty} V(\theta_i) M^{-1}_{\theta_i}\left(\Gamma_i\right)\delta_{\theta_i},$$

where L^{-1}_{n,θ_i}, $M^{-1}_{\theta_i}\left(\Gamma_i\right)$ are defined as in Theorem 8 and $V(\theta_i)$ are independent Dirichlet of random vectors with parameters $\gamma_1(\theta_i), \ldots, \gamma_n(\theta_i)$. Here B is the beta-Dirichlet process with parameters $(A_0, c, \gamma_1, \gamma_1, \ldots, \gamma_n)$. We refer the reader to the paper of Kim, James and Weibbach (2012) for the details.

4 Empirical Results: Comparison to Other Methods

Sampling from the beta process plays a central role in many applications. We refer the reader to the work of Paisley and Carin (2009) and Broderick, Jordan, and Pitman (2012). It is also required to simulate the beta-Dirichlet process (Kim, James and Weibbach, 2012). Several algorithms to sample from the beta process exist in the literature. In this section, we compare the new approximation of the beta process with the algorithm of Ferguson and Klass (1972), the algorithm of Damien, Laud, and Smith (1995), the algorithm of Wolpert and

Ickstadt (1998), the algorithm of Lee and Kim (2004) and the algorithm of Lee (2007). A summary of these algorithms is given in Appendix A.

In order to make comparisons between the preceding algorithms, we use equivalent settings for the parameters characterizing these algorithms (see Table 1). We consider the beta process with $c(t) = 2e^{-t}$ and $A_0(t) = t$, where $t \in [0, 1]$. We compute the absolute maximum difference between an approximate sample mean and the exact mean. See also Lee and Kim (2004) and Lee (2007) for similar comparisons. The exact mean of $A(t)$ in this example is $A_0(t) = t$; see Hjort (1990). We refer to this statistic by the maximum mean error. Specifically,

$$\text{maximum mean error} = \max_t |E[A_n(t)] - E[A(t)]| \quad = \quad \max_t |E[A_n(t)] - t|,$$

where $t = 0.1, 0.2, \ldots, 0.9, 1.0$ and A_n is an approximation of $A \sim BP(c(t) = 2e^{-t}, A_0(t) = t)$. Note that $E[A_n(x)]$ is approximated by obtaining the mean at $t = 0.1, 0.2, \ldots, 0.9, 1.0$ of 3000 i.i.d. sample paths from the approximated process A_n. Similarly, we compute the maximum standard deviation error between an approximate sample standard deviation (s.d.) and the exact standard deviation. The exact standard deviation of $A(t)$ is $\sqrt{t/3}$; see Hjort (1990). Thus,

$$\begin{aligned} \text{maximum s.d. error} &= \max_t |s.d\,[A_n(t)] - s.d.\,[A(t)]| \\ &= \max_t \left| s.d.\,[A_n(t)] - \sqrt{t/3} \right|. \end{aligned}$$

Table 1 depicts values of the maximum mean error, the maximum standard deviation error, and the corresponding computational time. Simulating the algorithm of Ferguson and Klass (1972) and the algorithm of Wolpert and Ickstadt (1998) is performed through relatively complex numerical methods, which are not appropriate for many users and time consuming (See Table 1). The R func-

tion "uniroot" is used to implement these two algorithms. The computational time is computed by applying the R function "System.Time". As seen in Table 1, the new algorithm has the smallest mean and standard deviation errors. Furthermore, it has a very reasonable computation time.

Table 1: This table reports the maximum mean error, the maximum standard deviation error, and the corresponding computation time. Here, FK, DSL, WI and LK stand for the algorithm of Ferguson and Klass (1972), the algorithm of Damien, Laud, and Smith (1995), the algorithm of Wolpert and Ickstadt (1998), and the algorithm of Lee and Kim (2004), respectively.

Algorithm	Parameters	max. mean error	max. s.d. error	Time
KL	$n = 200$	0.0192	0.1047	657.53
DSL	$m = n = 200$	0.0167	0.0145	90.53
WI	$M = 200$	0.0167	0.0884	406.10
LK	$\epsilon = 0.01$	0.0217	0.0239	0.29
Lee	$n = 200, \epsilon = 0.05$	0.0125	0.0522	1.25
New	$n = 200$	0.0069	0.0089	6.28

5 Conclusions

The vague convergence of random measures of the form (1) has been studied in this paper. An interesting application of the derived results includes deriving a finite sum representation that converges vaguely to the representation of Wolpert and Ickstadt (1998) of the beta process. This representation gives a simple yet efficient approach to approximate the beta process. We believe that the comprehensive study of metrizing random measures as in (1) and its strong association to various Bayesian nonparametric priors will add further useful

tools to the Bayesian nonparametric toolbox.

6 Acknowledgments

Research of the author is supported by the **Natural Sciences and Engineering Research Council of Canada (NSERC)**.

References

[1] Al-Labadi, L., and Abdelrazeq, I. (2016). On Functional Central Limit Theorems of Bayesian Nonparametric Priors. *Statistical Methods & Applications* (Springer). DOI:10.1007/s10260-016-0365-8.

[2] Al-Labadi, L., and Zarepour, M. (2013a). A Bayesian nonparametric goodness of fit test for right censored data based on approximate samples from the Beta-Stacy process. *Canadian Journal of Statistics*, 41, 3, 466-487.

[3] Al-Labadi, L. and Zarepour, M. (2013b). On asymptotic properties and almost sure approximation of the normalized inverse-Gaussian process. *Bayesian Analysis*, 8, 3, 553-568.

[4] Al-Labadi, L., and Zarepour, M. (2014a). Goodness of fit tests based on the distance between the Dirichlet process and its base measure. *Journal of Nonparametric Statistics*, 26, 2, 341-357.

[5] Al-Labadi, L., and Zarepour, M. (2014b). On simulations from the two-parameter Poisson-Dirichlet process and the normalized inverse-Gaussian pn mrocess. *Sankhyā A* , 76, 1, 158-176.

[6] Broderick, T, Jordan, M. I., and Pitman, J. (2012). Beta processes, stick-breaking, and power laws. *Bayesian Analysis*, **7**, 439-476.

[7] Damien, P., Laud, P. and Smith, A. F. M. (1995). Approximate random variate generation from infinitely divisible distributions with applications to Bayesian inference. *Journal of the Royal Statistical Society: Series B*, 57, 547-563.

[8] de Haan, L., and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer, New York.

[9] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 2, 209-230.

[10] Ferguson, T. S. and Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *Annals of Mathematical Statistics*, 43, 5, 1634-1643.

[11] Grandell, J. (1977). Point processes and random measures. *Advances in Applied Probability*, 9, 3, 502-526.

[12] Hjort, N. L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics*, 18, 3, 1259-1294.

[13] Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes. *Annals of Statistics*, 27, 2, 562-588.

[14] Kim, Y., James, L. and Weibbach, R. (2012). Bayesian analysis of multistate event history data: Beta-dirichlet process prior. *Biometrika*, 99, 1, 127-140.

[15] Lee, J. (2007). Sampling methods for neutral to the right processes. *Journal of Computational and Graphical Statistics*, 16, 3, 656-671.

[16] Lee, J. and Kim, Y. (2004). A new algorithm to generate beta processes. *Computational Statistics & Data Analysis*, 47, 3, 441-453.

[17] Lijoi, A., Mena, R. H. and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse Gaussian priors. *Journal of the American Statistical Association*, 100, 472, 1278-1291.

[18] Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer, Switzerland.

[19] Paisley, J. and Carin, L. (2009). Nonparametric factor analysis with beta process priors. In Proceedings of the International Conference on Machine learning (ICML).

[20] Phadia, E. G. (2013). *Prior Processes and Their Applications: Nonparametric Bayesian Estimation*. Springer-Verlag, Berlin.

[21] Pitman, J. and Yor, M. (1997). The two-parameter PoissonDirichlet distribution derived From a stable subordinator. *Annals of Probability*, 25, 2, 855-900.

[22] Walker, S. and Muliere, P. (1997). Beta-stacy processes and a generalisation of the polya-urn scheme. *Annals of Statistics*, 25, 4, 1762-1780.

[23] Wolpert, R. L., and Ickstadt, K. (1998). Simulation of Lévy random fields. In *Practical Nonparametric and Semiparametric Bayesian Statistics*. Eds. Day, D., Nuller, P., and Sinha, D., Springer, 2237-242.

[24] Zarepour, M., and Al-Labadi, L. (2012). On a rapid simulation of the Dirichlet process. *Statistics & Probability Letters*, 82, 5, 916-924.

A Other Sampling Algorithms

Below is a brief discussion of the algorithms considered in Section 4 of the present paper. We refer the reader to the original papers for more details.

Let $A \sim BP\left(c(\cdot), A_0(\cdot)\right)$ on $[0, t_0]$, where $t_0 > 0$ is fixed. We assume that $A_0(t_0) < \infty$.

• **Ferguson-Klass Algorithm:** The steps of the algorithm of Ferguson and Klass (1972) are:

1. Let $\Gamma_i = E_1 + \cdots + E_i$, where $(E_i)_{i \geq 1}$ are i.i.d. random variables with exponential distribution of mean 1.

2. For each $i \geq 1$, let J_i be the solution of $\Gamma_i = L_{t_0}\left(J_i\right)$, where $L_{t_0}\left(x\right) = L_{t_0}\left([x, 1)\right)$, $x > 0$ and the measure L_t is given by (2).

3. Generate i.i.d. random variables $(U_i)_{i \geq 1}$ from the uniform distribution on $[0, 1]$, independent of $(E_i)_{i \geq 1}$.

4. For $i \geq 1$, let θ_i be the solution of $U_i = n_{\theta_i}(J_i)$ in $[0, t_0]$ where

$$n_t(s) = \frac{\int_0^t c(z) s^{-1} (1-s)^{c(z)} dA_0(z)}{\int_0^{t_0} c(z) s^{-1} (1-s)^{c(z)} dA_0(z)} = \frac{\int_0^t c(z)(1-s)^{c(z)} dA_0(z)}{\int_0^{t_0} c(z)(1-s)^{c(z)} dA_0(z)}.$$

The process

$$A(t) = \sum_{i=1}^{\infty} J_i I(\theta_i \leq t) = \sum_{i=1}^{\infty} L_{t_0}^{-1}(\Gamma_i) I(\theta_i \leq t) \tag{6}$$

is a beta process with parameters $c(\cdot)$ and $A_0(\cdot)$. This series is an infinite series. In practice, we truncate this series and use the approximation

$$A_n(t) = \sum_{i=1}^{n} J_i I(\theta_i \leq t) = \sum_{i=1}^{n} L_{t_0}^{-1}(\Gamma_i) I(\theta_i \leq t).$$

• **Damien-Laud-Smith Algorithm:** Using the fact that the distributions of the increments of a nondecreasing Lévy process are infinitely divisible, Damien, Laud, and Smith (1995) derived an algorithm to generate approximations for infinitely divisible random variables and used it to generate the beta process.

First, the time interval $[0, t_0]$ is partitioned into small subintervals with end-points $0 = \theta_0 < \theta_1 < \ldots < \theta_m = t_0$. Let p_i denotes the increment of the process A in the interval $\Delta_i = (\theta_{i-1}, \theta_i]$, i.e. $p_i = A(\theta_i) - A(\theta_{i-1})$. The steps of the Damien-Laud-Smith algorithm for simulating the beta process are:

(1) Fix a relatively large positive integer n.

(2) Generate independent values z_{ij} from Π, where $\Pi(dz) = \eta(dz)/A_0(t_0)$ and $\eta([0,t]) = A_0(t)$, for $j = 1, \ldots, n$.

(3) Generate $x_{ij} \sim beta(1, c(z_{ij}))$, for $j = 1, \ldots, n$.

(4) Generate y_{ij}: $y_{ij}|x_{ij} \sim Poisson(\lambda_i n^{-1} x_{ij}^{-1})$, for $j = 1, \ldots, n$, where $\lambda_i = A_0(\theta_i) - A_0(\theta_{i-1})$.

(5) Set $p_{i,n} = \sum_{j=1}^{n} x_{ij} y_{ij}$. For large n, $p_{i,n}$ is an approximation of p_i.

(5) Set $A_n = \sum_{i=1}^{n} p_{i,n} \delta_{\theta_{(i)}}$.

Damien, Laud, and Smith (1995) showed that $A_n \xrightarrow{d} A$, as $n \to \infty$.

• **Wolpert-Ickstadt Algorithm:** The steps of the algorithm of Wolpert and Ickstadt (1998) are:

1. For $i = 1, 2, \ldots$, generate $\theta_i \overset{i.i.d.}{\sim} \Pi$, where $\Pi(dz) = \eta(dz)/A_0(t_0)$ and $\eta([0,t]) = A_0(t)$.

2. Let $\Gamma_i = E_1 + \cdots + E_i$, where $(E_i)_{i \geq 1}$ are i.i.d. random variables with exponential distribution of mean 1, independent of $(\theta_i)_{i \geq 1}$.

3. Define
$$M_z(x) = \int_x^\infty A_0(t_0) c(z) s^{-1} (1-s)^{c(z)-1} ds.$$

4. For each $i \geq 1$, solve the equation

$$M_{\theta_i}(J_i) = \Gamma_i$$

for J_i, where $\Gamma_i = E_1 + \cdots + E_i$, $(E_i)_{i\geq 1}$ are i.i.d. random variables with exponential distribution of mean 1 and independent of $(\theta_i)_{i\geq 1}$.

5. Set

$$A(t) = \sum_{i=1}^{\infty} M_{\theta_i}^{-1}(\Gamma_i) I\left\{\theta_i \leq t\right\}. \tag{7}$$

The process A in (7) is a beta process with parameters $c(\cdot)$ and $A_0(\cdot)$. This series is an infinite series. In practice, we truncate this series and use the approximation

$$A_n(t) = \sum_{i=1}^{n} M_{\theta_i}^{-1}(\Gamma_i) I\left\{\theta_i \leq t\right\}.$$

• **Lee-Kim Algorithm:** First the Lévy measure (2) of the beta process is approximated by

$$L_{t,\epsilon}(ds) = \left[\int_0^t \frac{c(s)}{\epsilon} b(s : \epsilon, c(z)) dA_0(z)\right] ds,$$

where

$$b(x : a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad \text{for } 0 < x < 1, a > 0, b > 0. \tag{8}$$

The steps of the algorithm of Lee and Kim (2004) for the beta process are:

(1) Fix a relatively small positive number ϵ.

(2) Generate the total number of jumps $n \sim Poisson\,(\lambda_\epsilon)$, where $\lambda_\epsilon = L_{t_0,\epsilon}\left((0,1)\right) = \epsilon^{-1} \int_0^{t_0} \int_0^1 c(z) b(s : \epsilon, c(z)) ds dA_0(z) = \epsilon^{-1} \int_0^{t_0} c(z) dA_0(z)$

(3) Generate the jump times $\theta_1, \ldots, \theta_n$ form the probability density function $dG_\epsilon/\lambda_\epsilon$, where $dG_\epsilon(z) = \epsilon^{-1} c(z) dA_0(z) I(0 \leq z \leq t_0)$.

(4) Let $\theta_{(1)} \leq \ldots \leq \theta_{(n)}$ be the corresponding order statistics of $\theta_1, \ldots, \theta_n$.

(5) Generate the jump sizes $p_1, \ldots, p_n : p_i|\theta_{(i)} \sim Beta(\epsilon, c(\theta_{(i)}))$.

(6) Set $A_\epsilon = \sum_{i=1}^n p_i \delta_{\theta_{(i)}}$.

Lee and Kim (2004) showed that $A_\epsilon \stackrel{d}{\to} A$, as $\epsilon \to 0$.

- **Lee Algorithm:** The steps of the algorithm of Lee (2007) are:

(1) Fix a relatively large positive integer n.

(2) For $i = 1, 2, \ldots, n$, generate $\theta_i \stackrel{i.i.d.}{\sim} \Pi$, where $\Pi(dz) = \eta(dz)/A_0(t_0)$ and $\eta([0, t]) = A_0(t)$.

(3) For $i = 1, \ldots, n$, generate $x_i \sim b(s : \epsilon, c(\theta_i))$, where $b(s : \epsilon, c(\theta_i))$ is defined in (8).

(4) For $i = 1, \ldots, n$, generate $y_i \sim Poisson\left(A_0(t_0)b(x_i : 1, c(\theta_i))/(nx_i b(x_i : \epsilon, c(\theta_i)))\right)$.

(5) Set $A_n = \sum_{i=1}^n x_i y_i \delta_{\theta_i}$.

Lee (2007) proved that, as $n \to \infty$, $A_n \stackrel{d}{\to} A$.